Entropy and mutual information

4^{th} November 2005

Definition 1. A *probability space* is a triple (S, B, P) on the domain S, which is a nonempty set called the *sample space*, where (S, B) is a measurable space, B is a Borel field of subsets of S, and P is a measure on S with the property that P(S) = 1 and, for all disjoint $E_i \in \mathrm{B}$,

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty}\right) = \sum_{i=1}^{\infty} \mathrm{P}(E_i)$$

In other words, P is a nonnegative function defined for all events $E_i \in \mathrm{B}$, and B measurable subsets of S. Further, a *random variable* X is a function mapping S into some set R, called the range of X. For convenience, we shall also use X to represent both the function and its own range, that is X is a function which maps S into X. If S is discrete and f is some real-valued function defined on S, then both X and f(X) are two different random variables, and the expectation of the latter is given by,

$$\mathrm{E}\left[\mathrm{f}(\mathrm{X})\right] = \sum_{x} \mathrm{p}(x)\mathrm{f}(x)$$

§

Definition 2. Let $\mathrm{p}(x)$ be the probability that $x \in \mathrm{X}$ occurs, similarly $\mathrm{p}(y)$ that $y \in \mathrm{Y}$ does-, while $\mathrm{p}(x, y)$ that both $x \in \mathrm{X}$ and $y \in \mathrm{Y}$ do occur. Then,

$$\mathrm{p}(x|y) = \frac{\mathrm{p}(x, y)}{\mathrm{p}(y)} \tag{1}$$

and

$$\mathrm{p}(y|x) = \frac{\mathrm{p}(x, y)}{\mathrm{p}(x)} \tag{2}$$

§

Definition 3. A *Markov chain* is a set of random variable X_t, where $t = 0, 1, \ldots$, such that,

$$\mathrm{P}\left(\mathrm{X}_t = j|\mathrm{X}_0 = i_0, \ldots, \mathrm{X}_{t-1} = i_{t-1}\right) = \mathrm{P}\left(\mathrm{X}_t = j|\mathrm{X}_{t-1} = i_{t-1}\right)$$

In other words, given the present state, the next state is conditionally independent of the past.

§

Definition 4. A subset $\mathrm{K} \subseteq \mathrm{E}^n$, where E^n is the Euclidean space of n dimensions, is called *convex* if the line segment joining any two points in K is contained in K. Let the two points be x_1 and x_2, then the line segment joining them together is $x = tx_1 + (1-t)x_2$, where $0 \le t \le 1$.

§

Definition 5. A point x is said to be a *convex combination* of points $x_1, \ldots, x_m$ if there exist nonnegative scalars $\alpha_1, \ldots, \alpha_m$ such that $\sum \alpha_i = 1$ and $\sum \alpha_i x_i = x$. The set of all convex combinations of x_i, $i = 1, \ldots, m$, is called the *convex hull* of $\{x_i\}$.

§

Definition 6. Let f be a real-valued function, and let K be a convex subset of the domain of f. Then f is said to be *convex cup* if, for every $x_1, x_2 \in \mathrm{K}$ and $0 \le t \le 1$,

$$\mathrm{f}(tx_1 + (1-t)x_2) \le t\mathrm{f}(x_1) + (1-t)\mathrm{f}(x_2) \tag{3}$$

It is said to be *strictly convex cup* if strict inequality holds in Equation 3 whenever $x + 1 \neq x_2$. Similarly, f is said to be *convex cap* if,

$$\mathrm{f}(tx_1 + (1-t)x_2) \ge t\mathrm{f}(x_1) + (1-t)\mathrm{f}(x_2) \tag{4}$$

that is to say, if $-f$ is convex cup. It is said to be *strictly convex cap* if strict inequality holds in Equation 4 whenever $x_1 \neq x_2$. Convex cap is also known as *concave*. Geometrically speaking, f is convex cup if and only if all its chords lie above or on the graph of f, and f is concave if and only if all its chords lie below or on the graph of the same.

§

Definition 7. Let K be some interval in E^1, and let $\mathrm{F}(x)$ be a probability distribution concentrated on K such that $\mathrm{P}(\mathrm{X} \leq x) = \mathrm{F}(x)$. Then, if the expectation E(X) exists, and if $\mathrm{f}(x)$ is a convex cup function, then,

$$\mathrm{E}(\mathrm{f}(\mathrm{X})) \geq \mathrm{f}(\mathrm{E}(\mathrm{X})) \tag{5}$$

If f is strictly convex cup, then strict inequality holds in Equation 5. Similarly, if f is convex cap, then,

$$\mathrm{E}(\mathrm{f}(\mathrm{X})) \leq \mathrm{f}(\mathrm{E}(\mathrm{X}) \tag{6}$$

If f is strictly convex cap, then strict inequality holds in Equation 6.

§

Example 1. Suppose that in Definition 7 there is a mass distribution placed on the graph of f, then Equation 5 says that the overall centre of mass will lie above or on the graph, while Equation 6 says that it will lie below it.

Entropy is a measure of uncertainty of many events as a single value. We derive it from Axiom's 1 and 2.

Axiom 1. If the events are all equally likely, then the uncertainty function $\mathrm{H}\left(\frac{1}{m},\ldots,\frac{1}{m}\right)$ is monotonously increasing with m.

§

Axiom 2. If $\{E_1^1,\ldots,E_m^1\}$ and $\{E_1^2,\ldots,E_n^2\}$ are statistically independent sets of equally likely disjoint events, then the uncertainty of the sets of events $\{E_i \cap E_j; i = 1,\ldots,m; j = 1,\ldots,n\}$ is

$$\mathrm{H}\left(\frac{1}{mn},\ldots,\frac{1}{mn}\right) = \mathrm{H}\left(\frac{1}{m},\ldots,\frac{1}{m}\right) + \mathrm{H}\left(\frac{1}{n},\ldots,\frac{1}{n}\right)$$

That is to say, $\mathrm{h}(mn) = \mathrm{h}(m) + \mathrm{h}(n)$, where $\mathrm{h}(m) = \mathrm{H}\left(\frac{1}{m},\ldots,\frac{1}{m}\right)$.

§

Definition 8. Let the set of m possible disjoint events be $\mathrm{E} = \{E_1,\ldots,E_m\}$. We call an *apriori probability* of E_i, $\mathrm{p}(E_i)$, where $1 \leq i \leq m$ and $\sum_{i=1}^m \mathrm{p}(E_i) = 1$. The *uncertainty function* or the *entropy function*, $\mathrm{H}(\mathrm{p}(1),\ldots,\mathrm{p}(m))$ obeys Axiom's 1 and 2.

§

The entropy of a random variable x gives a measure of the amount of *information* obtained from an observation of x. It also represents the *randomness* of x and our *uncertainty* about x. The less probable an event is, the more information we receive when it occurs.

Theorem 1. The entropy of a set of m equally likely events is $\mathrm{h}(m) = \lambda \log_c m$, where λ is a positive constant and $c > 1$.

Proof. Proving Theorem 1 amounts to proving that Axiom's 1 and 2 are satisfied if and only if $\mathrm{h}(m) = \lambda \log_c m$. The two axioms say that $\mathrm{h}(m)$ is monotonously increasing in m and

$$\mathrm{h}(m) = \mathrm{h}(m) + \mathrm{h}(n) \tag{7}$$

According to Equation 7, if $m = n = 1$, then $\mathrm{h}(1) = \mathrm{h}(1) + \mathrm{h}(1)$, which implies that $\mathrm{h}(1) = 0$. From this together with both axioms above, $\mathrm{h}(m) = \lambda \log_c m$ is sufficient as a solution.

Next, we must prove that this solution is necessarily the only solution. Let a, b and c be positive integers, and $a, b, c > 1$. Then there exists a unique integer d such that

$$c^d \leq a^b < c^{d+1} \tag{8}$$

From Equation 8 it follows that,

$$d \log c \leq b \log a < (d+1) \log c$$

and therefore,

$$\frac{d}{b} \leq \frac{\log a}{\log c} < \frac{d+1}{b} \tag{9}$$

Since $\mathrm{h}(m)$ is monotonously increasing, from Equation 8 we have,

$$\mathrm{h}(c^d) \leq \mathrm{h}(a^b) < \mathrm{h}(c^{d+1})$$

Then from Equation 7, $d\mathrm{h}(c) \leq b\mathrm{h}(a) < (d+1)\mathrm{h}(c)$. And since $\mathrm{h}(m)$ is monotonously increasing,

$$\frac{d}{b} \leq \frac{\mathrm{h}(a)}{\mathrm{h}(c)} < \frac{d+1}{b} \tag{10}$$

From Equation's 9 and 10 it follows that,

$$\left|\frac{\log a}{\log c} - \frac{\mathrm{h}(a)}{\mathrm{h}(c)}\right| < \frac{1}{b}$$

And, since b is arbitrary positive integer,

$$\begin{aligned} \frac{\mathrm{h}(a)}{\mathrm{h}(c)} &= \frac{\log a}{\log c} \\ \frac{\mathrm{h}(a)}{\log a} &= \frac{\mathrm{h}(c)}{\log c} \end{aligned}$$

Since a and c are arbitrary,

$$\frac{\mathrm{h}(a)}{\log a} = \lambda = \frac{\mathrm{h}(c)}{\log c}$$

Therefore, necessarily $\mathrm{h}(m) = \lambda \log_c m$ is the only solution. ¶

Axiom 3. The total uncertainty of events does not depend on the method of indication. §

Axiom 4. The uncertainty measure is a continuous function with regard to the probabilities within it. §

Example 2. Let a set E of m disjoint events be $\{E_1, \ldots, E_m\}$. Let j_i, $i = 0, \ldots, n$, be integers and $0 = j_0 \leq j_1 < j_2 \cdots < j_n = m$, and E be divided into n sets of events, namely,

$$\begin{aligned} \mathrm{G}_1 &= \{E_1, \ldots, E_{j_1}\} \\ \mathrm{G}_2 &= \{E_{j_1+1}, \ldots, E_{j_2}\} \\ &\vdots \\ \mathrm{G}_n &= \{E_{j_{n-1}+1}, \ldots, E_m\} \end{aligned}$$

If we indicate firstly the group, and then the event within that group, then the uncertainty becomes,

$$\mathrm{H}(\mathrm{p}(E_1), \ldots, \mathrm{p}(E_m)) = \mathrm{H}(\mathrm{p}(\mathrm{G}_1), \ldots, \mathrm{p}(\mathrm{G}_n)) + \sum_{i=1}^{n} \mathrm{p}(\mathrm{G}_i)\mathrm{H}(\mathrm{p}(E_{j_{i-1}+1}|\mathrm{G}_i), \ldots, \mathrm{p}(\mathrm{E}_{j_i}|\mathrm{G}_i)) \tag{11}$$

The grouping axiom, Axiom 3, lets us express the uncertainty when all the event probabilities are rational. By grouping equally likely events together and then consider each of the groups as a single event, it gives us the ability to deal with events which are not equally likely. Example 3 gives an example how this is done. Then Axiom 4 extends Axiom 3 to cover also irrational probabilities, and Equation 12 is the result.

Example 3. As in Example 2, let a set of disjoint events be $\mathrm{E} = \{E_1, \ldots, E_m\}$, and let $\mathrm{p}(E_i) = \frac{1}{m}$, $i = 1, \ldots, m$. Also, let the groups of events $\mathrm{G}_1, \ldots, \mathrm{G}_n$ be defined the same way therein. Let n_k be the number of events in G_k. Then $n_k = j_k - j_{k-1}$ and $\mathrm{p}(\mathrm{G}_k) = \frac{n_k}{m}$, for $k = 1, \ldots, n$, and also $\mathrm{p}(E_i|\mathrm{G}_k) = \frac{1}{n_k}$, for $j_{k-1} < i \leq j_k$. Then Equation 11 yields,

$$\mathrm{h}(m) = \mathrm{H}(\mathrm{p}(\mathrm{G}_1), \ldots, \mathrm{p}(\mathrm{G}_n)) + \sum_{i=1}^{n} \mathrm{p}(\mathrm{G}_i)\mathrm{h}(n_i)$$

And since from Theorem 1, $\mathrm{h}(m) = \lambda \log_c m$, we have,

$$\begin{aligned} \mathrm{H}(\mathrm{p}(\mathrm{G}_1), \ldots, \mathrm{p}(\mathrm{G}_n)) &= -\sum_{i=1}^{n} \mathrm{p}(\mathrm{G}_i)(\mathrm{h}(n_i) - \mathrm{h}(m)) \\ &= -\sum_{i=1}^{n} \mathrm{p}(\mathrm{G}_i) \left(\lambda \log \frac{n_i}{m}\right) \\ &= -\lambda \left(\sum_{i=1}^{n} \mathrm{p}(\mathrm{G}_i) \log \mathrm{p}(\mathrm{G}_i)\right) \ (12) \end{aligned}$$

Example 4. From $\mathrm{h}(m) = \lambda \log_c m$, if we let $\lambda = \log_b c$, then $\mathrm{h}(m) = \log_b m$. In other words, the scale factor λ can be absorbed in the base of the logarithm.

Theorem 2. Let $\{p_1, \ldots, p_m\}$ be a set of probabilities such that $\sum_{i=1}^{m} p_i = 1$. Then, †

$$\mathrm{H}(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i \tag{13}$$

Proof. This is the results from Example's 2 and 3, and the scale factor λ disappears in a manner similar to that shown by Example 4. ¶

Example 5. If the base of the logarithm in Equation 13 is 2, the unit of the entropy is *bit*. On the other hand if this base is e, that is to say, if we use natural logarithms, then the uncertainty has the unit of *nat*. From this, one may see that one nat is equal to $\log_2 e$ bits, which is approximately 1.443 bits. The term *bit* comes from *binary digit*, the term *nat* from *natural digit*.

Definition 9 explains what is meant by *conditional entropy*. Starting from Equation 14, which is an equation for conditional entropy when y is given, we obtain the overall conditional entropy in Theorem 3. For any pair of sets X and Y given, H(X|Y) gives the amount of uncertainty remaining about X after Y has been observed.

Definition 9. The *conditional entropy* of X, given some $y \in \mathrm{Y}$, is,

$$\mathrm{H}(\mathrm{X}|y) = -\sum_{x} \mathrm{p}(x|y) \log \mathrm{p}(x|y) \tag{14}$$

Then the conditional entropy H(X|Y) is the expectation, or average value, of $\mathrm{H}(\mathrm{X}|y)$ over the range Y. In other words,

$$\mathrm{H}(\mathrm{X}|\mathrm{Y}) = \sum_{y} \mathrm{p}(y)\mathrm{H}(\mathrm{X}|y) \tag{15}$$

§

Theorem 3. The conditional entropy is,

$$\mathrm{H}(\mathrm{X}|\mathrm{Y}) = -\sum_{x,y} \mathrm{p}(x,y) \log \mathrm{p}(x|y)$$

† Some times the entropy function is defined instead by $\mathrm{H}(p_1, \ldots, p_m) = \sum_{i=1}^{m} p_i \log \frac{1}{p_i}$, but this is obviously the same as our Equation 13 since $\log x^{-1} = -\log x$.

Proof. Putting the equation of conditional entropy when y is given, Equation 14, into the overall conditional entropy equation, Equation 15, we get,

$$\begin{aligned}\mathrm{H}(\mathrm{X}|\mathrm{Y}) &= \sum_y \mathrm{p}(y)\mathrm{H}(\mathrm{X}|y)\\ &= -\sum_y \mathrm{p}(y)\sum_x \mathrm{p}(x|y)\log \mathrm{p}(x|y)\end{aligned}$$

Then from Equation 1 of Definition 2, $\mathrm{p}(y)\mathrm{p}(x|y) = \mathrm{p}(x,y)$, and so,

$$\mathrm{H}(\mathrm{X}|\mathrm{Y}) = -\sum_{x,y} \mathrm{p}(x,y)\log \mathrm{p}(x|y)$$

¶

Theorem 4. Let X, Y and Z be discrete random variables. For each $z \in \mathrm{Z}$, let $\mathrm{E}(z) = \sum_{x,y}\mathrm{p}(y)\mathrm{p}(z|x,y)$. Then,

$$\mathrm{H}(\mathrm{X}|\mathrm{Y}) \le \mathrm{H}(\mathrm{Z}) + \mathrm{E}(\log \mathrm{E})$$

Proof.

$$\begin{aligned}\mathrm{H}(\mathrm{X}|\mathrm{Y}) &= -\mathrm{E}\left[\log \mathrm{p}(x|y)\right]\\ &= -\sum_{x,y,z} \mathrm{p}(x,y,z)\log \mathrm{p}(x|y)\\ &= -\sum_z \mathrm{p}(z)\sum_{x,y}\frac{\mathrm{p}(x,y,z)}{\mathrm{p}(z)}\log \mathrm{p}(x|y)\end{aligned}$$

Because

$$\frac{\mathrm{p}(x,y,z)}{\mathrm{p}(z)} = \mathrm{p}(x,y|z)$$

is a probability distribution, that is a convex cap function, we may apply Equation 6, namely Jensen's inequality for convex cap, from Definition 7. Hence,

$$\begin{aligned}\mathrm{H}(\mathrm{X}|\mathrm{Y}) &\le \sum_z \mathrm{p}(z)\log\left(\frac{1}{\mathrm{p}(z)}\sum_{x,y}\frac{\mathrm{p}(x,y,z)}{\mathrm{p}(x|y)}\right)\\ &= \sum_z \mathrm{p}(z)\log\frac{1}{\mathrm{p}(z)} + \sum_z \mathrm{p}(z)\log\sum_{x,y}\frac{\mathrm{p}(x,y,z)}{\mathrm{p}(x|y)}\end{aligned}$$

But,

$$\frac{\mathrm{p}(x,y,z)}{\mathrm{p}(x|y)} = \frac{\mathrm{p}(x,y,z)\mathrm{p}(y)}{\mathrm{p}(x,y)} = \mathrm{p}(y)\mathrm{p}(z|x,y)$$

hence the statement above is proved. ¶

Corollary 4[1]. Let X adn Y be random variables each of which takes values in the set $\{x_1,\ldots,x_r\}$. Let $\mathrm{P}_e = \mathrm{P}(\mathrm{X}\ne\mathrm{Y})$. Then,

$$\mathrm{H}(\mathrm{X}|\mathrm{Y}) \le \mathrm{H}(\mathrm{P}_e) + \mathrm{P}_e\log(r-1)$$

Proof. From Theorem 4, let $\mathrm{Z} = 0$ if $\mathrm{X} = \mathrm{Y}$, and let $\mathrm{Z} = 1$ if $\mathrm{X} \ne \mathrm{Y}$. Then $\mathrm{E}(0) = 1$ and $\mathrm{E}(1) = r-1$. ¶

Theorem 5. The maximum uncertainty occurs when the events are equiprobable.

Proof. Since,

$$\begin{aligned}\mathrm{H}\left(\frac{1}{m},\ldots,\frac{1}{m}\right) - \mathrm{H}(p_1,\ldots,p_m) &= \log_b m + \sum_{i=1}^m p_i\log_b p_i\\ &= \log_b e\sum_{i=1}^m p_i \ln mp_i\\ &\ge \log_b e\sum_{i=1}^m p_i\left(1-\frac{1}{mp_i}\right) = 0\end{aligned}$$

it being the case that $\ln\frac{1}{x} \ge 1-x$. Therefore $\mathrm{H}(p_1,\ldots,p_m)$ is maximised when $p_i = \frac{1}{m}$, for all $i = 1,\ldots,m$. ¶

Example 6. Figure 1 shows that $\ln x \leq x - 1$, while Figure 2 shows that such inequality does not exist when the logarithm in question is of base 10.

Figure 1 *Plots of* $\ln x$ *and* $x - 1$*, which show that* $\ln x \leq x - 1$.

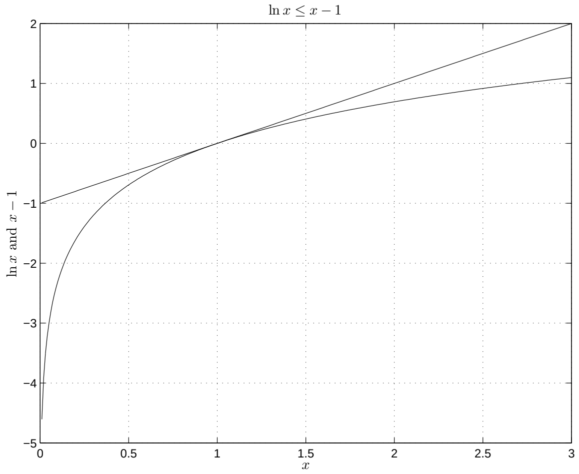
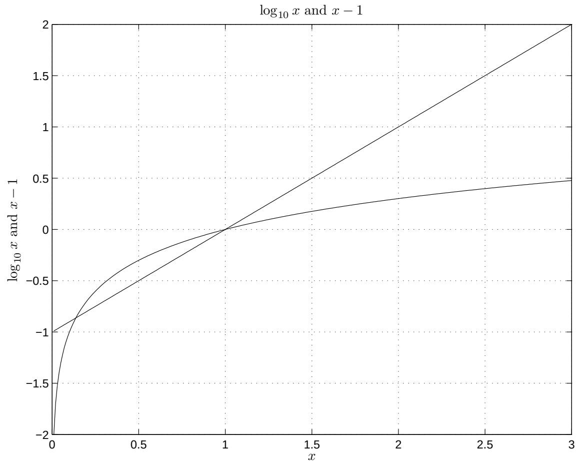


Figure 2 *Graphs of* $y = \log x$ *and* $y = x - 1$*, which show that the latter is no bound for the values of the former.*



Example 7. Figure 3 confirms for us how $\ln \frac{1}{x} \geq 1 - x$, whereas Figure 4 tells us that this is the case for $\log \frac{1}{x}$.

Figure 3 *Plots showing* $\ln \frac{1}{x}$ *and* $1 - x$*, which show that* $\ln \frac{1}{x} \geq 1 - x$.

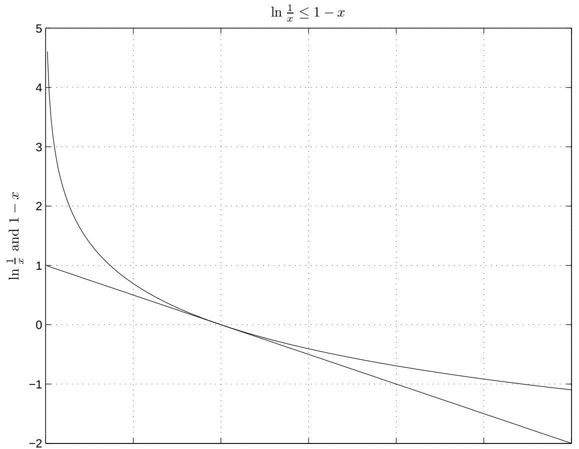
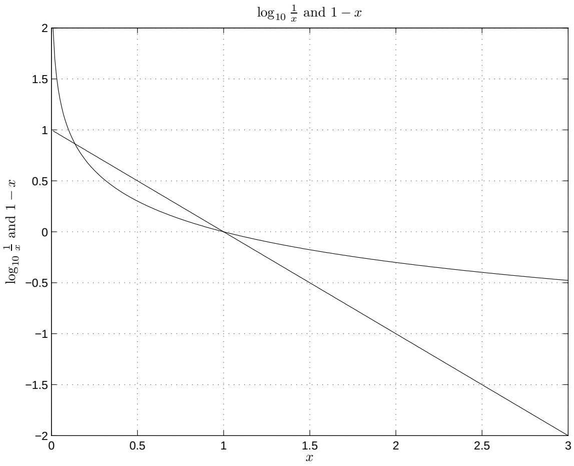


Figure 4 *Graphs showing* $y = \log \frac{1}{x}$ *and* $y = 1 - x$*, from which it is clear the latter gives no bounds for the former.*

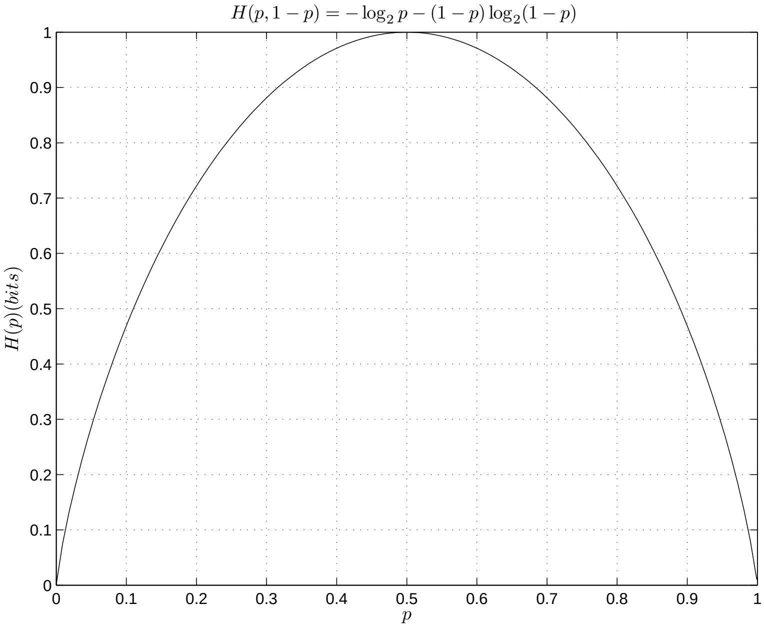


Example 8. Consider two events with probabilities p and $1-p$. The entropy function is then,

$$\mathrm{H}(p, 1-p) = -p \log p - (1-p)\log(1-p)$$

Whenever the occurrence of either event become certainty, the entropy function would become zero. Mathematically we see that $\lim_{p\to 0} p\log p = 0$ and $\lim_{p\to 1} p\log p = 0$. Figure 5 shows a plot of the values of the entropy function for two events. Base-2 logarithm is used here.

Figure 5 *The entropy function of two events with probabilities p and $1-p$.*



Definition 10. The *mutual information* is $\mathrm{I}(\mathrm{X};\mathrm{Y}) = \mathrm{H}(\mathrm{X}) - \mathrm{H}(\mathrm{X}|\mathrm{Y})$. It represents the information provided about X by Y.

§

Example 9. Alternatively, the mutual information may take the following form, *cf* Definition 2,

$$\begin{aligned}\mathrm{I}(\mathrm{X};\mathrm{Y}) &= \sum_{x,y} \mathrm{p}(x,y)\log\frac{\mathrm{p}(x|y)}{\mathrm{p}(x)}\\ &= \sum_{x,y} \mathrm{p}(x,y)\log\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)}\\ &= \sum_{x,y} \mathrm{p}(x,y)\log\frac{\mathrm{p}(y|x)}{\mathrm{p}(y)}\end{aligned}$$

That is to say, I(X; Y) is the average taken over the X, Y sample space of the random variable I($x; y$0 such that,

$$\mathrm{I}(x;y) = \log\frac{\mathrm{p}(x|y)}{\mathrm{p}(x)} = \log\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)} = \log\frac{\mathrm{p}(y|x)}{\mathrm{p}(y)}$$

Theorem 6. For any discrete random variables X and Y, $\mathrm{I(X;Y)} \geq 0$. Moreover, $\mathrm{I(X;Y)} = 0$ if and only if X and Y are independent.

Proof. From one of our formulae for the mutual information and from Jensen's inequality,

$$\begin{aligned}\mathrm{I(X;Y)} &= -\sum_{x,y}\log\frac{\mathrm{p}(x)\mathrm{p}(y)}{\mathrm{p}(x,y)}\\ &\geq \log\sum_{x,y}\mathrm{p}(x)\mathrm{p}(y) = \log 1 = 0\end{aligned}$$

Furthermore, the equality sign holds if and only if $\mathrm{p}(x)\mathrm{p}(y) = \mathrm{p}(x,y)$ for all x and y, that is to say, when X and Y are independent of each other. ¶

Example 10. From our formulae of the mutual information, we may see that,

$$\mathrm{I(X;Y)} = \mathrm{I(Y;X)}$$

and

$$\mathrm{I(X;Y)} = \mathrm{H(Y)} - \mathrm{H(Y|X)}$$

Also,

$$\mathrm{I(X;Y)} = \sum_{x,y}\mathrm{p}(x,y)\log\frac{1}{\mathrm{p}(x,y)}$$

Definition 11. Let X, Y and Z be three random variables. Then the mutual information I(X, Y; Z) is given by,

$$\mathrm{I(X,Y;Z)} = \mathrm{E}\left(\log\frac{\mathrm{p}(z|x,y)}{\mathrm{p}(z)}\right) = \sum_{x,y,z}\mathrm{p}(x,y,z)\log\frac{\mathrm{p}(z|x,y)}{\mathrm{p}(z)}$$

This mutual information is the amount of information X and Y provide about Z.

§

Theorem 7. Let X, Y and Z be three random variables. Then we have $\mathrm{I(X,Y;Z)} \geq \mathrm{I(Y;Z)}$, where the equality holds if and only if $\mathrm{p}(z|x,y) = \mathrm{p}(z|y)$ for all (x,y,z) such that $\mathrm{p}(x,y,z) > 0$.

Proof.

$$\begin{aligned}\mathrm{I(Y;Z)} - \mathrm{I(X,Y;Z)} &= \mathrm{E}\left(\log\frac{\mathrm{p}(z|y)}{\mathrm{p}(z)} - \log\frac{\mathrm{p}(z|x,y)}{\mathrm{p}(z)}\right)\\ &= \mathrm{E}\left(\log\frac{\mathrm{p}(z|y)}{\mathrm{p}(z|x,y)}\right)\\ &= \sum_{x,y,z}\mathrm{p}(x,y,z)\log\frac{\mathrm{p}(z|y)}{\mathrm{p}(z|x,y)}\end{aligned}$$

Then using Jensen's inequality, we have,

$$\begin{aligned}\mathrm{I(Y;Z)} - \mathrm{I(X,Y;Z)} &\leq \log\sum_{x,y,z}\mathrm{p}(x,y,z)\frac{\mathrm{p}(z|y)}{\mathrm{p}(z|x,y)}\\ &= \log\sum_{x,y,z}\mathrm{p}(x,y)\mathrm{p}(z|y) = \log 1 = 0\end{aligned}$$

¶

Theorem 8. Let (X, Y, Z) be a Markov chain. Then,

$$I(X; Z) \leq \begin{cases} I(X; Y) \\ I(Y; Z) \end{cases}$$

Proof. From Theorem 7, $I(X; Z) \leq I(X, Y; Z)$. Because (X, Y, Z) is a Markov chain, $I(X, Y; Z) = I(Y; Z)$. Therefore $I(X; Z) \leq I(Y; Z)$. Next, since (X, Y, Z) is a Markov chain, (Z, Y, X) is also a Markov chain. Hence $I(X; Z) \leq I(X; Y)$. ¶

Bibliography

Solomon W Golomb, Robert E Peile and Robert A Scholtz. *Basic concepts in information theory and coding, The adventures of Secret Agent 00111.* Plenum Press, New York, 1994

Robert J McEliece *The theory of information and coding.* Addison-Wesley, 1977